数据分析

数据处理内容：

1. 重新选择肉类供应链——对中文分类字段“product\_classification”进行选择“肉类”， 根据excel“禽肉类产品分类表”重新圈出肉类产品，并分为鲜肉和制品两类；

2. 对违规原因重新分类（划分为五大类）

风险类型参考表（122种adulterant\_english）占比分析【我会先制作风险类型对应表】（表格）

3. 对检测单位进行识别（属于供应链的哪个环节,数据段已分类，需进一步合并）

禽畜肉类供应链

注：选择2013（或2014）至2018（或2019）一个完整的时间跨度进行后续分析，尽量保证前后一致。

1. 抽检数据概述：不同年份不同省份畜禽肉类与果蔬类的抽检数据量及总体合格率；**（输出表格）**

2. 各类风险(更新后的风险类别)总体分布情况；存在不合格但未标注的情况(说明不合格且未标注原因的检测数据量),按照已标注的数据/总数计算。**（输出表格）**

3. 畜禽肉类中各类风险分布情况同上，不合格条目及不合格率**（输出表格）**

4. 时间跨度内所有数据以及肉类产品数据来看各供应链节点的检测力度(即检测数量) 、合格率。**（输出表格）**

注：结合production year和announcement year 解决空值问题。

5. 肉类：不同风险类型（更新后的风险类型）在供应链节点（区分鲜肉和制品）上的分布**（输出表格）**

6. 对禽肉类供应链节点上的风险指数进行建模（已有模型待完善）观察其分布，并求出不同节点风险指数随时间的变化**（输出表格）**

溯源：

1. 重点：不同节点上的风险可能来源，并估计各风险源引入各节点的可能（需要一份风险与其对应的来源的表格）【我这周和老师商量后发过来对应表再操作】

2. 对供应链上高风险厂商进行标记，并查看是否进行了二次检测，检测合格率如何，与全国检测合格率对标比较。根据year（或者半年）分析风险厂商和取样点的合格率和检测频率(表格重点)

3. 相关性分析：生产商和取样点：抽检频次与抽检合格率的关系（引出根据风险指数安排动态监管的概念）（检测次数、检测规模与合格率的关系）【根据公司划分企业规模？产品类别，这周确定】

省份：

（以浙江省作为分析对象，与全国平均水平进行对比、与四川辽宁、北京等地进行对比）

1. 对省份所属地区进行合格率分析(属于东部、西部、中部)以观察区域差异;(省份划分为各区域的表格我会出)**（输出表格）**

2. 对四川、辽宁、北京、西藏、宁夏、上海、浙江等省份的每年和总体的检测量和风险分布情况做分析**（输出表格）**

3. 一份地图展示全国各地的风险分布情况（高风险、低风险）【大小表示】形式待确定**（输出热力图或散点图）**

4. 相关性分析：分析省份差异（生产商和抽样处的距离、肉类产品（生鲜肉类而非冷冻肉制品）、生产时间、检测级别与合格率的关系）尝试回归，如果无法回归，则描述其检测力度。【之后确定】